O'REILLY®

LIVE ONLINE TRAINING

Web Scraping in 60 Minutes

https://resources.oreilly.com/binderhub/advanced-web-scraping

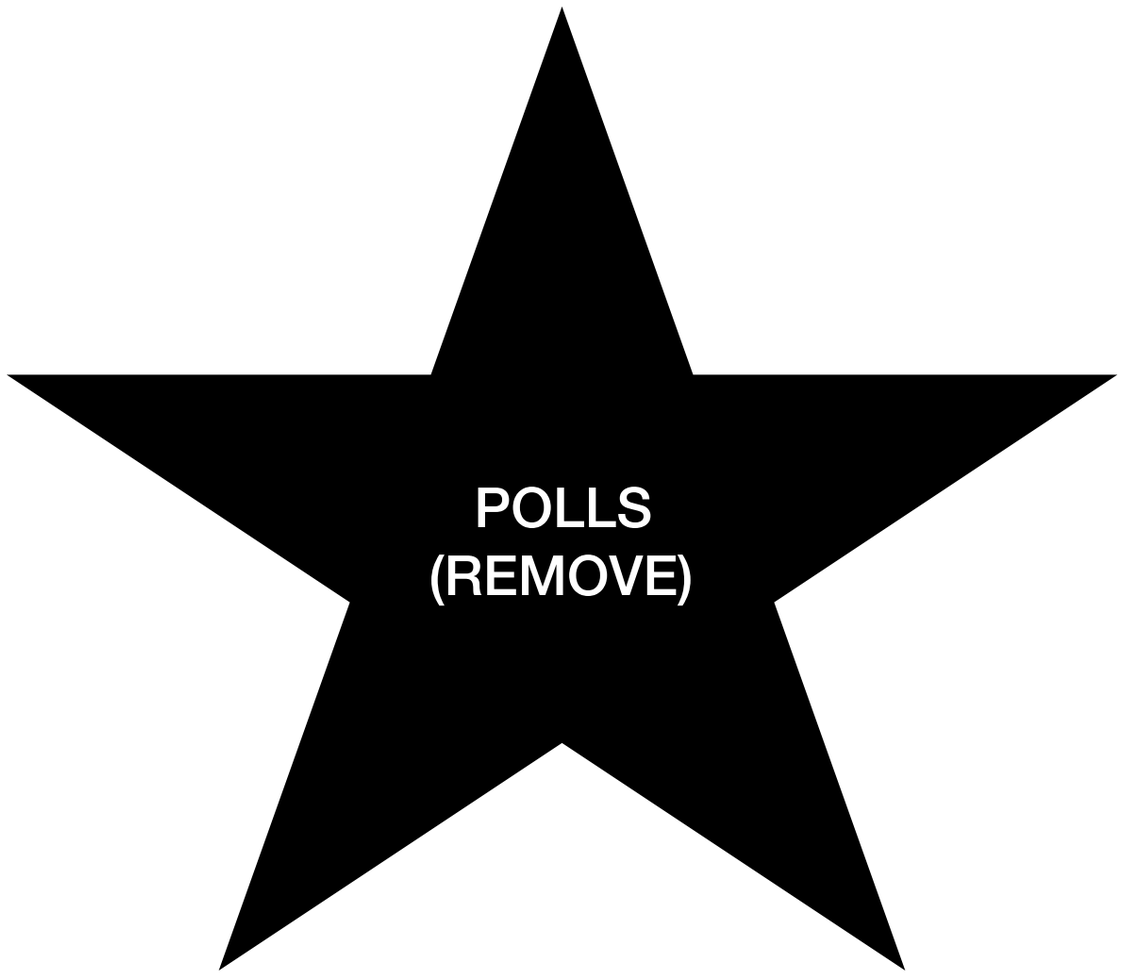
April 2, 2020

10:00am – 11:00am EDT

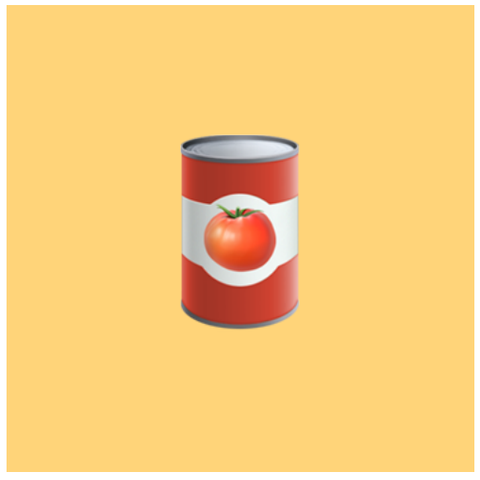
1

2

3



- Poll: How many websites have you scraped before? {0, 1, 10, 100+}
- Poll: Is your interest in web scraping professional or personal? {professional, personal}



https://gazpacho.xyz/

gazpacho is a web scraping library.
It replaces requests and
BeautifulSoup for *most* projects.

https://gazpacho.xyz/


y tho

X0,000 LOC
not on GitHub
lxml dependancy
learning curve
15 years old

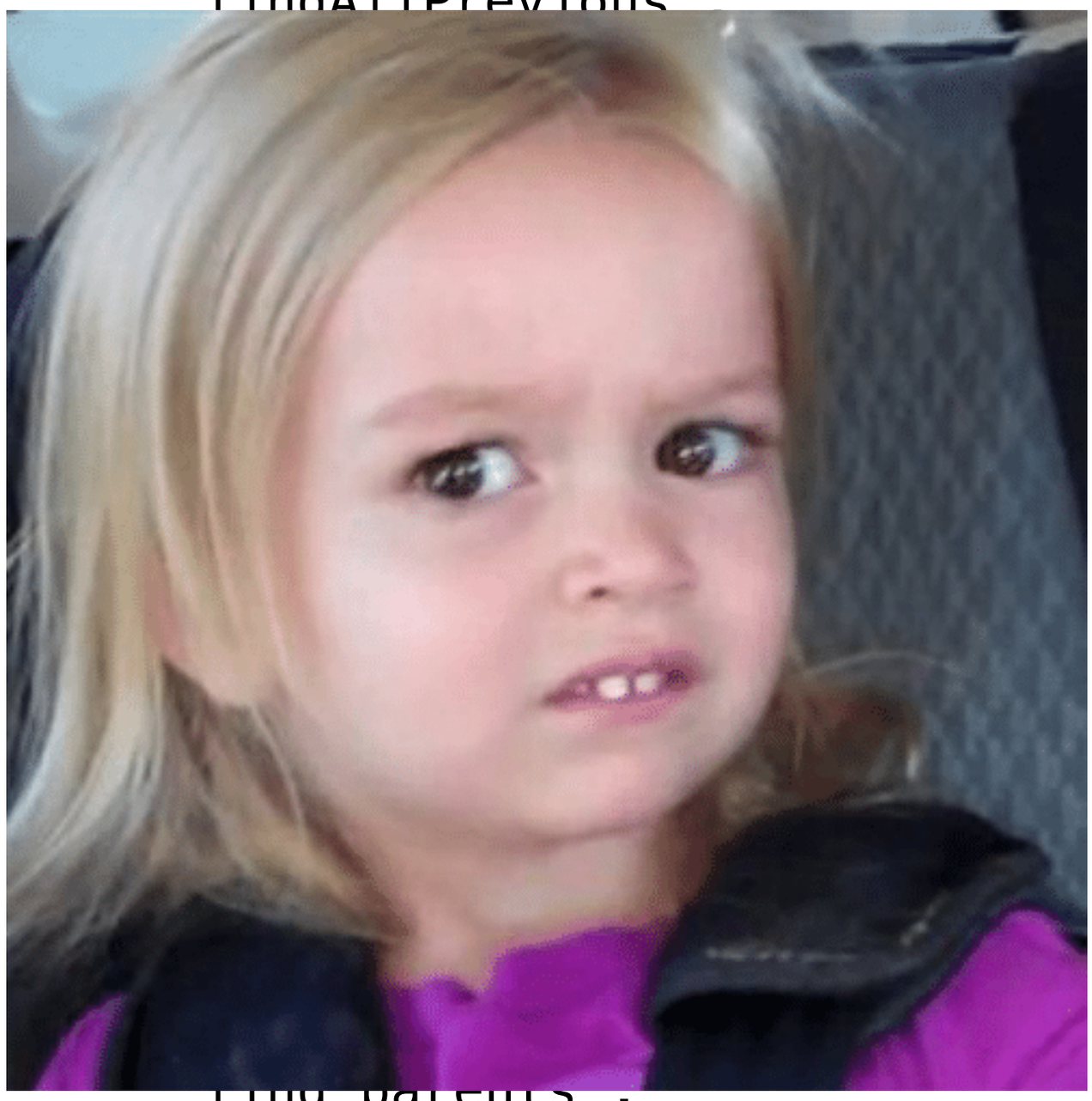
```
[‘find',
 'findAll',
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious',
 'findChild',
 'findChildren',
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious',
 'findChild',
 'findChildren',
 'findNext',
 'findNextSibling',
 'findNextSiblings',
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious',
 'findChild',
 'findChildren',
 'findNext',
 'findNextSibling',
 'findNextSiblings',
 'findParent',
 'findParents',
 'findPrevious',
 'findPreviousSibling',
 'findPreviousSiblings',
 'find_all',
 'find_all_next',
 'find_all_previous',
 'find_next',
 'find_next_sibling',
 'find_next_siblings',
 'find_parent',
 'find_parents',
 'find_previous',
 'find_previous_sibling',
 'find_previous_siblings']
```

```
[‘find',
 'findAll',
 'findAllNext',
 'findAllPrevious'
```



```
 'find_parents',
 'find_previous',
 'find_previous_sibling',
 'find_previous_siblings']
```

find

✅ 0 dependancies
✅ one method (find)
✅ X00 LOC
✅ fast (30% & 300%)
✅ 1 package

https://gazpacho.xyz/

```
pip install gazpacho
```

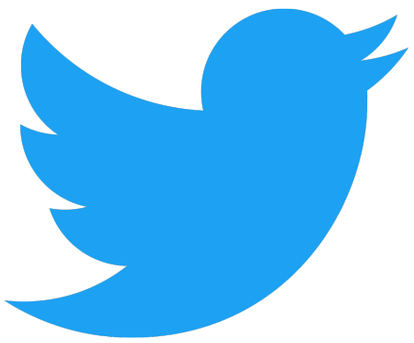

jupyter

Q&A

That's all Folks!

github.com/maxhumber/mummify

github.com/maxhumber/gif



twitter.com/maxhumber

www.linkedin.com/in/maxhumber